ECON3500 Lab 8: Instrumental variables

It's our final lab of the semester!

Materials

- `voucher.dta`
- Do-file template `econ3500_lab_template.do`

Objectives

Today we're going to work with `voucher.dta`, a dataset of student performance from Rouse (1998). She measures the impact of private school vouchers on student achievement.

By the end of this lab, you should be able to complete the following tasks in Stata:

- Estimate instrumental variable specifications and interpret them.
- Output regression results using `outreg2`

Why instrumental variables?

In Labs 6 and 7, we dealt with **endogeneity** — situations where our key independent variable is correlated with the error term, usually because of omitted variables, measurement error, or reverse causality. Fixed effects (Lab 7) solve this when the problem comes from time-invariant confounders.

Instrumental variables (IV) offer another approach: find a variable (the "instrument") that affects Y *only through* X. This instrument provides a source of exogenous variation in X that we can use to estimate the causal effect. The two requirements for a valid instrument are: 1. **Relevance**: The instrument must be correlated with the endogenous variable (X). 2. **Exclusion restriction**: The instrument must affect Y only through X (not directly).

Data context

The data come from an evaluation of the Milwaukee Parental Choice Program, which randomly offered school vouchers to students via a lottery. The final measure of student performance is `mnce`, their math test score in 1994 (after up to four years in a private school). We also have baseline performance: their math test score in 1990 (`mnce90`). The variable `choiceyrs` is the number of years actually enrolled in a private school, and `selectyrs` is the number of years a student was *selected* (via lottery) to receive a voucher.

The lottery creates a natural instrument: being *selected* for a voucher (which is random) affects the number of years *enrolled* in a private school, but shouldn't directly affect test scores through any other channel.

Variables we'll use

variable	meaning	notes
`mnce`	math score in 1994	outcome variable
`mnce90`	math score in 1990	baseline performance
`choiceyrs`	years enrolled in a choice school	endogenous variable

variable	meaning	notes
`selectyrs`	years selected to receive a voucher	instrument
`choiceyrs1`–`choiceyrs4`	dummies for years in choice school	used in Q9
`selectyrs1`–`selectyrs4`	dummies for years selected for voucher	used in Q9
`black`	Black indicator	
`hispanic`	Hispanic indicator	
`female`	female indicator	

Key commands

command	description
`ivregress 2sls y (x = z) controls, robust`	IV regression using two-stage least squares
`ivregress 2sls y (x = z) controls, robust first`	Same, reporting first-stage results
`predict yhat, xb`	Generate predicted values from the previous regression
`testparm varname`	Test significance of a coefficient (F-statistic)
`outreg2 using file.xls, replace`	Export regression results to Excel (first column)
`outreg2 using file.xls, append`	Add a column to an existing results table

Conducting IV regressions with `ivregress`

General form:

```
ivregress estimator depvar [varlist1] (varlist2 = varlist_iv) [if] [in] [weight] [, options]
```

- `estimator` is where we will type `2sls`
- `depvar` is your dependent variable
- You can include other explanatory variables before or after the parentheses, `[varlist1]`
- In the parentheses, write your endogenous (x) then your instrument (z) — these can be lists!
- The rest of it is just as you're used to

Example:

To estimate the following two-stage least squares equation:

$$rent = \beta_0 + \beta_1 \widehat{hsngval} + \beta_2 pcturban + u$$

where $\widehat{hsngval}$ is predicted from the following first-stage equation

$$hsngval = \alpha_0 + \alpha_1 faminc + \alpha_2 pcturban + v$$

```
webuse hsng2

ivregress 2sls rent (hsngval = faminc) pcturban, robust
```

You can add `, first` to report the first-stage results:

```
ivregress 2sls rent (hsngval = faminc) pcturban, robust first
```

Outputting your results with `outreg2`

We are very good at reading raw Stata output. But raw Stata output has no place in our papers. How do we make it pretty? There are lots of ways, including `putexcel`, which lets you create customizable Excel

tables with your outputs (good for descriptive statistics), and `estout`, which does the same thing but is more regression oriented.

Personally, I like `outreg2`, because it's easy to set up and use. So that's what we'll use!

Installation required: `outreg2` is a user-created package, which means you have to install it first:

```
ssc install outreg2
```

You only need to do this **once** per computer. If you get an error that `outreg2` is already installed, that's fine — just keep going.

You'll run `outreg2` after estimating a regression. It takes your results and saves them to a table. You can run it multiple times and generate columns of results within the same Excel sheet, which is pretty handy! The general format of `outreg2` is this:

```
// You can copy and paste this into Stata, and it should work!
// Note that it will save to your working directory

sysuse auto, clear

// Specification 1
regress mpg foreign weight headroom trunk length turn displacement
outreg2 using myfile.xls, replace

// Specification 2  (add on)
regress mpg foreign weight headroom trunk length turn displacement, robust
outreg2 using myfile.xls, append
```

You can customize with lots of options! (See `help outreg2`, or check out these resources)

What sort of things?

- Export directly to Word
  - `outreg2 using myfile, word replace`
- Add notes
  - `outreg2 using myfile, addnote(Dummy variables not shown)`
- Report only some variables
  - `outreg2 using myfile, keep(mpg foreign)`
- Modify number of decimal places
  - `outreg2 using myfile, dec(5)`
- You can use a loop to make a whole set of columns!

An example:

```
sysuse auto, clear
local r "replace"
  forval num = 1/5 {
      regress mpg weight headroom if rep78 == `num'
      sum mpg if rep78 == `num'
      local mean = `r(mean)'
      outreg2 using myfile.xls, `r' keep(headroom) title("Sample Graph") nocons addtext("Rep78", `num'

  local r "append"
  }
```

Workflow overview

1. Load `voucher.dta` and start your log file.

2. Explore the data (`summarize`, `describe`).
3. Estimate OLS regressions (naive estimates).
4. Run the first stage and check instrument relevance.
5. Estimate IV models (by hand and with `ivregress`).
6. Compare OLS and IV results.
7. Create a summary table with `outreg2`.

Lab 8 Worksheet

What do I submit?

- Your written-up answers to exercise questions (1)–(10). This can be typed or written out then scanned (or photographed), in any reasonable format.
- The do-file you created that runs this analysis
- A log file that contains the results from this exercise.
- **A table with your regression results** (six columns, from `outreg2`). Include this with your written answers.

Use robust standard errors in all regressions.

Questions

1. In your do-file, start a log and open `voucher.dta`.
2. Summarize your data. Of the 990 students in the sample, how many were never awarded a voucher? How many had a voucher for all four years? How many actually attended a choice school for four years?

Hint: `tab selectyrs` and `tab choiceyrs` will show you the distribution.

3. Predict the relationship between choice school attendance and math scores by regressing math scores `mnce` (dependent variable) on number of years enrolled in a choice school `choiceyrs` (independent variable). What do you find? Is this what you expect? What happens if you add in the variables `black`, `hispanic`, and `female`? Write your results in equation form.
4. Why might `choiceyrs` be endogenous? Explain.
5. Now, estimate a regression of `choiceyrs` (dependent variable) on `selectyrs` (independent variable), including race/ethnicity and gender controls. Why is this a reasonable choice of an instrument? What is the F-statistic on `selectyrs`?

Hint: Use `testparm selectyrs` after the regression to get the F-statistic. A rule of thumb is that the F-statistic should be at least 10 for the instrument to be considered strong enough.

6. Based on the previous regression, use the `predict` command to generate a predicted $\widehat{choiceyrs}$. Estimate the regression of `mnce` on $\widehat{choiceyrs}$, including race/ethnicity and gender controls. Write the estimated equation. How does your result compare to your OLS estimate?

Reminder: The `predict` command generates fitted values from the most recently estimated regression. Run it immediately after the Q5 regression — before running anything else:

```
predict choiceyrs_hat, xb
```

Then use `choiceyrs_hat` as your independent variable in the second-stage regression.

7. Re-estimate a regression of `mnce` (dependent variable) on `choiceyrs` (independent variable) using `selectyrs` as an instrument for `choiceyrs`. This time, estimate the equation in one command line using `ivregress 2sls`. How do your results change, if at all?

Example syntax:

```
ivregress 2sls mnce (choiceyrs = selectyrs) black hispanic female, robust
```

Important: The *coefficients* from Q6 and Q7 should be the same, but the *standard errors* will differ. That's because the manual approach (Q6) doesn't correctly account for the fact that $\widehat{choiceyrs}$ is a generated regressor. `ivregress` adjusts the standard errors automatically — always use it in practice.

8. Repeat your IV analysis, but this time include a control for baseline achievement by adding `mnce90`. Write the results in equation form below. Do you find these results convincing? Explain.

Heads up: `mnce90` is missing for many students — your sample will drop from 990 to about 328 observations. This is expected. Think about what it means for your results.

9. We can also use multiple instruments for multiple endogenous variables. The variables `choiceyrs1`, `choiceyrs2`, etc. are dummy variables indicating the different number of years a student could have been in a choice school. Similarly, `selectyrs1`, `selectyrs2`, etc. have a similar definition, but for being selected from the lottery.

   Here, `choiceyrs1` = 1 if the student attended a choice school for exactly 1 year, `choiceyrs2` = 1 for exactly 2 years, and so on. The `selectyrs1`–`selectyrs4` variables are defined analogously for lottery selection.

   Estimate the following equation using IV:

$$mnce = \beta_0 + \beta_1 choiceyrs_1 + \beta_2 choiceyrs_2 + \beta_3 choiceyrs_3 + \beta_4 choiceyrs_4 + \beta_5 black + \beta_6 hispanic + \beta_7 female + \beta_8 mnce90 + u$$

Hint: Put all the endogenous variables on the left of the `=` and all the instruments on the right:

```
ivregress 2sls mnce ///
    (choiceyrs1 choiceyrs2 choiceyrs3 choiceyrs4 = ///
     selectyrs1 selectyrs2 selectyrs3 selectyrs4) ///
    black hispanic female mnce90, robust
```

10. Finally, go back through your regressions in your do-file. After each regression (there should be six: OLS without controls, OLS with controls, IV by hand, IV using `ivregress`, IV with `mnce90`, and IV with multiple instruments), add a line of code to output the results to a Word or Excel file using `outreg2`.

    Include a table with your results with your submission — there should be six columns in one table.

Hint: Use `replace` for the first regression and `append` for each subsequent one:

```
regress mnce choiceyrs, robust
outreg2 using lab8_results.xls, replace

regress mnce choiceyrs black hispanic female, robust
outreg2 using lab8_results.xls, append

// ... continue for remaining regressions
```

Submission checklist - Answers to questions (1)–(10) - Do-file with comments for each question - Log file that matches your do-file commands - `outreg2` table (six columns) - Make sure your do-file includes `log close` at the end

References: Rouse, Cecilia Elena (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *The Quarterly Journal of Economics* 113(2), 553-602.